Write up for “Relational similarity and the non-independence of features in similarity judgments” experimental simulations: statistics

Ruairidh McLennan Battleday
PhD COS
battleday@princeton.edu

1 Introduction

Five experiments are presented in Goldstone, Medin, and Gentner (1991) that examine the relationship between attributional and relational features in similarity judgements. We are interested in replicating the effect in Experiment 3, a similarity rating task that presents results averaged over subjects and six picture sets.

1.1 Materials and procedure

Participants were presented with pairs of stimuli (each composed of sets of objects), and made similarity comparisons between each stimulus ($\{A, B, C, D\}$) and the target ($\{T\}$), of the form “On a scale from 1 to 9, how similar are X and Y?”.

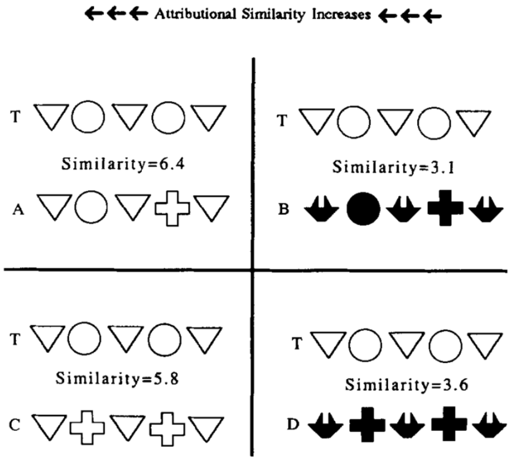

FIG. 4. Sample stimuli from Experiment 3. Feature independence is violated if subjects choose *D* as more similar to *T* than *B*, and choose *A* as more similar to *T* than *C*.

Figure 1: “On a scale from 1 to 9, how similar are $\{A, B, C, D\}$ and T?”

Participants were assessed over six picture groups (set 2 shown in Figure 1). The objects varied in the features SHAPE and SHADE, and the relations used were SAME-SHAPE and SAME-SHADING. Attributional similarity was varied by changing the shape or shading of one or more of the objects such that no relation was disturbed.

1.2 Behavioural data

Average results for each picture set are presented in Figure 2. Weak support for MAX is found if $(D - B) - (C - A)$ is positive. The authors report that “the mean similarity ratings, collapsing over the six picture groups, show the same trends, yielding a value of .6 for (D - B) - (C - A); this value is significantly greater than zero ($df = 5$, $t = 3.448$, $p < .02$)”. They also report that two of the individual stimulus sets have values significantly greater than zero (sets 2 and 3), but do not report the statistics; they also do not report whether these are significant under *post-hoc* correction for multiple testing, or whether they conducted any.

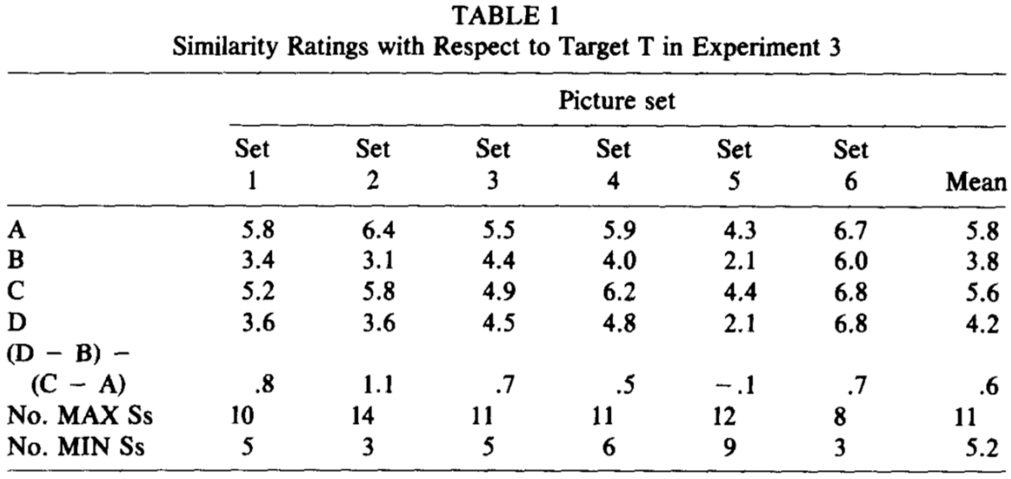

TABLE 1
Similarity Ratings with Respect to Target T in Experiment 3

	Picture set						
	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Mean
A	5.8	6.4	5.5	5.9	4.3	6.7	5.8
B	3.4	3.1	4.4	4.0	2.1	6.0	3.8
C	5.2	5.8	4.9	6.2	4.4	6.8	5.6
D	3.6	3.6	4.5	4.8	2.1	6.8	4.2
(D − B) − (C − A)	.8	1.1	.7	.5	−.1	.7	.6
No. MAX Ss	10	14	11	11	12	8	11
No. MIN Ss	5	3	5	6	9	3	5.2

Figure 2: Data from Exp. 3.

We can try to derive the effective sample size needed to replicate these effects with a power analysis, using set 3 (above) as a guideline of the minimum significant effect size for individual stimulus sets. On the basis of the reported statistic, the authors used a one-sample two-tailed t-test to test whether the mean difference differed from zero. That is, they assume the test statistic under the null hypothesis follows a Student’s t distribution with $n - 1$ degrees of freedom (formulae from Wikipedia):

$$t_{N-1} = \frac{\bar{X} - \mu_0}{S/\sqrt{N}} \tag{1}$$

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{N-1} \tag{2}$$

$$p(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu \cdot \pi} \cdot \Gamma(\frac{\nu}{2})}(1 + \frac{t^2}{\nu})^{-\frac{\nu+1}{2}}, \tag{3}$$

where $\nu = N - 1$ is the number of degrees of freedom.

Our goal is to calculate the sample size needed to replicate this effect given the significance level ($\alpha = 0.05$), as well as a power of 0.8 ($\beta = 0.2$). We know the mean for set 3 (0.7), as well as original sample size; however, we do not have access to individual-level data, and so cannot calculate the standard deviation.

1.2.1 Normal assumption

As Student’s t distribution is known to converge to the Normal with increasing degrees of freedom, we could estimate the effective sample size based on a Normal assumption. Recall that the formula for the related Z-statistic is given as follows:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}} \sim \text{Normal}(0, 1). \tag{4}$$

We can arrive at an estimate for N by considering a two-sided power calculation for z-tests. Recall the following consideration of power (`https://cran.r-project.org/web/packages/distributions3/vignettes/one-sample-z-test.html`):

$$p(\text{reject } H_0|\mu = \mu_A) = p(\bar{X} > \mu_0 + Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{N}}|\mu = \mu_A) + p(\bar{X} < \mu_0 + Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{N}}|\mu = \mu_A), \tag{5}$$

where μ_A is our alternative hypothesis, at $\bar{X}$ by construction. As $\bar{X} \sim \text{Normal}(\mu_A, \frac{\sigma^2}{N})$ by assumption, we can calculate these probabilities using simple rearrangements and the transformation given by the standard Normal distribution:

$$\begin{aligned} p(\bar{X} > \mu_0 + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{N}}|\mu = \mu_A) &= p\Big(\frac{\bar{X} - \mu_A}{\sigma/\sqrt{N}} > \frac{\mu_0 + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{N}} - \mu_A}{\sigma/\sqrt{N}}\Big), \qquad (6) \\ &= p\Big(Z > \frac{\mu_0 - \mu_A}{\sigma/\sqrt{N}} + z_{1-\frac{\alpha}{2}}\Big). \qquad (7) \end{aligned}$$

Similarly,

$$p(\bar{X} < \mu_0 + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{N}}|\mu = \mu_A) = p\Big(Z < \frac{\mu_0 - \mu_A}{\sigma/\sqrt{N}} + z_{\frac{\alpha}{2}}\Big). \tag{8}$$

Power at significance level α becomes as follows:

$$\text{Power} = p\Big(Z < \frac{\mu_0 - \mu_A}{\sigma/\sqrt{N}} + z_{\frac{\alpha}{2}}\Big) + p\Big(Z > \frac{\mu_0 - \mu_A}{\sigma/\sqrt{N}} + z_{1-\frac{\alpha}{2}}\Big). \tag{9}$$

Noting that only one of these terms contributes significantly if the null hypothesis has indeed been rejected (the parameter for alternative hypothesis can only more extreme in one direction), we can use algebra to recover N, as follows:

$$\begin{aligned} p\Big(Z < \frac{\mu_0 - \mu_A}{\sigma/\sqrt{N}} + z_{\frac{\alpha}{2}}\Big) &= 1 - \beta; \qquad (10) \\ \therefore z_\beta &= \frac{\mu_0 - \mu_A}{\sigma/\sqrt{N}} + z_{\frac{\alpha}{2}} \qquad (11) \\ \therefore N &= \Big(\frac{\sigma(z_\beta - z_{1-\alpha/2})}{\mu_0 - \mu_A}\Big)^2. \qquad (12) \end{aligned}$$

Using $\alpha = 0.05$, $\beta = 0.2$, $\mu_0 = 0$, and $\mu_A = 0.7$, the effect found in the previous paper, we have the following approximation for N:

$$N = \Big(\frac{\sigma \cdot (1.96 + 0.842)}{-0.7}\Big)^2. \tag{13}$$

Although we do not know σ, we can use our pilot data to derive an empirical plug-in estimator ($\hat{\sigma} = 1.701$). With all of these estimates at hand, we find $N \approx 47$.

Finally, we can plot the values we obtained during the pilot, to see whether normality is a reasonable assumption. In this case, it does not seem so.

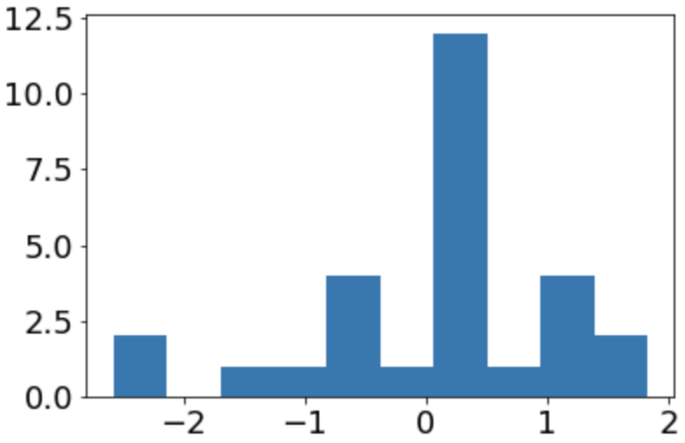

Figure 3: Data from pilot (D-B) - (C-A), z-scored.

1.2.2 t distribution

We could follow the same process using the t distribution directly. However, this runs into problems, as we shall see. Recall, the term S is:

$$S^2 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N-1} \tag{14}$$

This time, the relevant derivation is as follows (again WLOG restricting analysis to one tail):

$$p(\bar{X} < \mu_0 + t^{(n-1)}_{\alpha/2} \cdot S/\sqrt{N} | \mu = \mu_A) = p\Big(\frac{\bar{X} - \mu_A}{S/\sqrt{N}} < \frac{\mu_0 + t_{\alpha/2} \cdot \frac{S}{\sqrt{N}} - \mu_A}{S/\sqrt{N}}\Big), \tag{15}$$

$$= p\Big(T < \frac{\mu_0 - \mu_A}{S/\sqrt{N}} + t_{\alpha/2}\Big) \tag{16}$$

$$\therefore t^{(N-1)}_{\beta} = \frac{\mu_0 - \mu_A}{\sigma/\sqrt{N-1}} + t^{(N-1)}_{\alpha/2} \tag{17}$$

$$\therefore N = \Big(\frac{\sigma(t^{(N-1)}_{\beta} + t^{(N-1)}_{1-\alpha/2})}{\mu_0 - \mu_A}\Big)^2 + 1, \tag{18}$$

where we made use of the fact that $S/\sqrt{N} = \sigma/\sqrt{N-1}$.

The problem that immediately arises is that we need to know N to derive the degrees of freedom for the relevant t distribution, but we are trying to use that distribution to derive N in the first place. Some "authors" recommend beginning with the approximate N given by the Normal analysis above, and iterating until convergence (`https://stats.stackexchange.com/questions/146412/method-to-determine-the-sample-size-for-a-one-sample-t-test`).

Substituting $N = 47$, and the values above, we have:

$$\text{“}N\text{”} = \Big(\frac{1.701 \cdot (0.862 + 2.101)}{-0.7}\Big)^2 + 1, \tag{19}$$

$$= 50. \tag{20}$$

Obviously this is a discrepancy. However, we have convergence with one more iteration:

$$\text{DOF} = 49; N = 50. \tag{21}$$

We could also use the N from the last study to provide the initial degrees of freedom ($N = 29$). The convergence is as follows:

Obviously this is a discrepancy. However, we have convergence with one more iteration:

$$\text{DOF} = 29; N = 51; \tag{22}$$
$$\text{DOF} = 50; N = 50; \tag{23}$$
$$\text{DOF} = 49; N = 50. \tag{24}$$

1.2.3 Simulation

Another approach is to construct an empirical estimate of the distributions under different values of N using stochastic simulation, and test the significance and power of our effect size (e.g., following `https://nickch-k.github.io/EconometricsSlides/Week_08/Power_Simulations.html`).

That is, we sample $\{X_i\}_{i=1}^N \sim \text{Normal}(\mu_A, \hat{\sigma}^2)$, and derive the sufficient empirical statistics for our test (in this case, $(\tilde{\mu} = \bar{X}, \tilde{\sigma}); \tilde{S} = \sqrt{\frac{\sum_i (X_i - \bar{X})^2}{N-1}}$). We need to calculate the p-value and power of these data, and take the first N where the average of many simulations is in the range we want to find.

Recall, the formula for deriving the p-value of a two-sided one sample t-test is as follows:

$$t_{obs} = \frac{\tilde{\mu} - \mu_0}{\tilde{S}/\sqrt{N}}, \tag{25}$$
$$= \frac{\tilde{\mu} - \mu_0}{\tilde{\sigma}/\sqrt{N-1}}, \tag{26}$$
$$\text{DOF} = N - 1, \tag{27}$$

and we look up $p_{val} \equiv p(|T| \geq |t_{obs}|) = 1 - p(T \leq t_{obs}) + p(T \leq -t_{obs})$ using a t-table / inverse cumulative density function (function).

And, the power can be found by the following:

$$t_{obs} = \frac{\tilde{\mu} - \mu_0}{\tilde{\sigma}/\sqrt{N-1}}, \tag{28}$$
$$\text{DOF} = N - 1, \tag{29}$$
$$\text{Power} = p\Big(T < \frac{\mu_0 - \tilde{\mu}}{\tilde{S}/\sqrt{N}} + t_{\alpha/2}^{(\text{DOF})}\Big) + p\Big(T > \frac{\mu_0 - \tilde{\mu}}{\tilde{S}/\sqrt{N}} + t_{1-\alpha/2}^{(\text{DOF})}\Big) \tag{30}$$
$$= p\Big(T < \frac{\mu_0 - \tilde{\mu}}{\tilde{\sigma}/\sqrt{N-1}} + t_{\alpha/2}^{(\text{DOF})}\Big) + p\Big(T > \frac{\mu_0 - \tilde{\mu}}{\tilde{\sigma}/\sqrt{N-1}} + t_{1-\alpha/2}^{(\text{DOF})}\Big) \tag{31}$$
$$= p\Big(T < \frac{\mu_0 - \tilde{\mu}}{\tilde{\sigma}/\sqrt{N-1}} + t_{\alpha/2}^{(\text{DOF})}\Big) + \Big[1 - p\Big(T < \frac{\mu_0 - \tilde{\mu}}{\tilde{\sigma}/\sqrt{N-1}} + t_{1-\alpha/2}^{(\text{DOF})}\Big)\Big] \tag{32}$$

Running $M = 1000$ trials, and averaging across significance and power for each N, we obtain a Monte Carlo simulation estimate of $N \leftarrow 60$ (see Figure 4).

References

Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the nonindependence of features in similarity judgments. *Cognitive psychology*, *23*(2), 222–262.

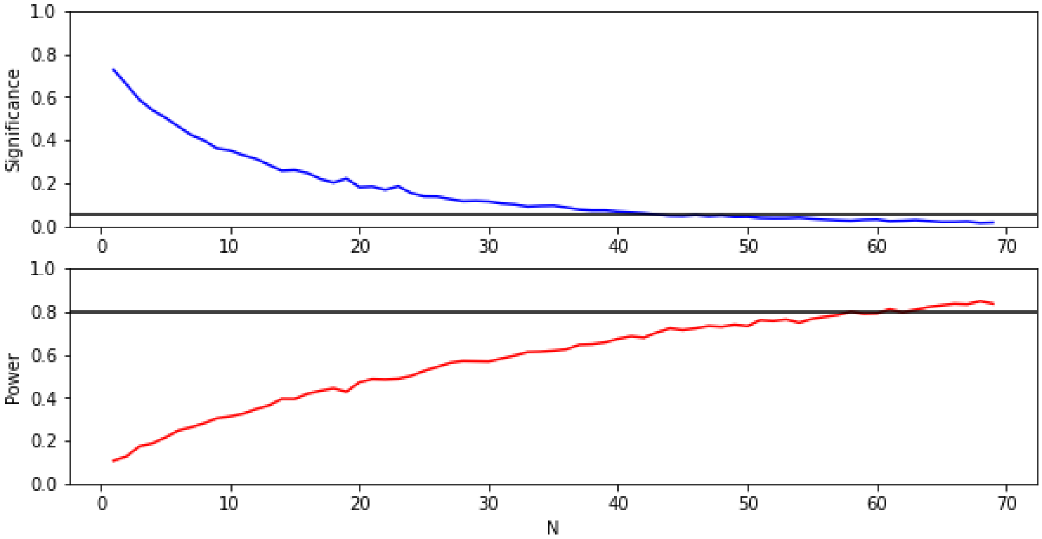

Figure 4: Simulation analysis of significance and power with N and the smallest effect size reported significant for a subset of stimuli in original study (set 3).